SIPP *Survey of Income and Program Participation*

Linking Core Wave, Topical Module, and Full Panel Files

This section describes reasons and procedures for linking files, including suggestions for handling nonmatches.

Reasons for Linking Files

Often, a single SIPP data file will not contain all the information needed for a project. In those cases, analysts may need to merge data from another file or link two or more files. For example, analysts often link SIPP files for the following reasons:

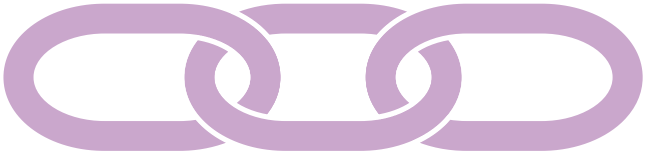
- Data for a single calendar reference month are often contained on two different core wave files.
- In the pre-1996 Panel files, data covering a single calendar year are often on files from two or even three different panels.
- Analysts may need to merge topical module data with core wave data.
- Analysts may need to link core wave files for a longitudinal analysis if the full panel file has not been released or if the variables of interest are not available in the longitudinal file (for pre-1996 files).

Procedures for Linking Files

In this tutorial section, and in Chapter 13 of the *SIPP Users*' *Guide*, procedures for linking person records across files are described. Procedures for linking households or families are problematic when working with longitudinal data—because unit composition changes over time—and are therefore not discussed.

Three Basic Steps

To link files, analysts need to:



1. Create data extracts from each file to be linked.
2. Sort the files in common order by using identified variables as match keys.
3. Merge the files.

Depending on the planned analysis and software used, analysts choose to create final files either in person-month format, reflecting the 1990 and later core wave files, or in person-record format.

Six Types of Merges

SIPP users commonly merge files in the following ways:

1. Within a core wave file
2. Two or more core wave files
3. Core wave and full panel files
4. Two or more topical module files
5. Topical module and core wave files
6. Topical module and full panel files

Information about the ID variables needed for the six types of merges is provided in Chapter 13 of the *SIPP Users*' *Guide*.

Descriptions of the Six Types of Merges

Merges Within a Core Wave File

Core wave files have one record per person per month. Linking within a core wave file transforms the files into a single wide record per person—the format used for core wave files before the 1990 Panel.

Chapter 13 of the *SIPP Users*' *Guide* describes two approaches for this linking process. Programmers using third-generation languages such as FORTRAN and PL/1 use one approach. Programmers using fourth-generation languages such as SAS and SPSS typically use the second approach. tip

SIPP*tip*

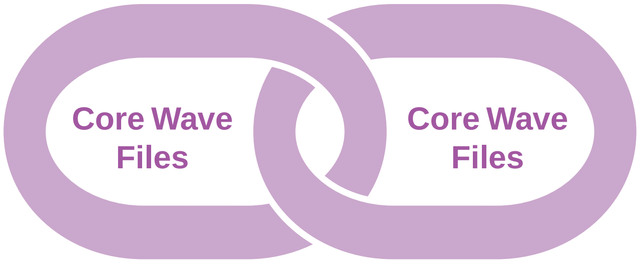
Chapter 13 of the SIPP Users' Guide *contains sample SAS code for changing core wave files from person-month format to person-record format.*

Merging Two or More Core Wave Files

There are two reasons to link two or more core wave files:

1. To create an analysis file with more than 4 months of information for each person

2. As a step in merging core wave data with data from another file type



To create a final-analysis file in person-month format from two or more waves, the analyst simply needs to sort and interleave the files. Refer to Chapter 13 of the *SIPP Users' Guide* for the necessary variables that will ensure a proper sort. To create files in person-record format with just one record per person, analysts first need to interleave files to create the person-month-format file. Analysts can then apply procedures for merging within a core wave file.

Effects of Editing and Imputation. Analysts should be careful when creating their own longitudinal databases from core wave files in the pre-1996 panels. All edits and imputations in a wave were independent of those used in other waves; thus, data across waves may be inconsistent. For basic demographic information, it is generally safe to assume that the most recent data are correct. tip

SIPP*tip*

New edit and imputation procedures that make use of prior wave data were used in the 1996 Panel to improve data consistency. Logical inconsistencies will still exist in the 1996 Panel files among reported items that were not longitudinally edited (basic demographic characteristics were longitudinally edited).

Weights. Analysts should note that the sample weights included on the core wave files are calendar month specific. These weights may not be appropriate for longitudinal analyses with linked core wave files.

Merging Core Wave and Full Panel Files

This procedure is not used very often because the two files contain the same information for the most part. However, some core information appears only on the core wave files, making it necessary at times to merge the core wave and full panel files.

To link data from the two file types, analysts should do the following:

1. Create data extracts from the core wave and full panel files.

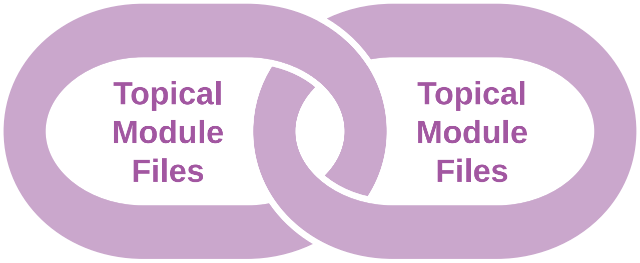
2. Put the extracts into the same format.
3. Sort the extracts in the same order.
4. Merge the extracts, creating the final file.

Chapter 13 of the *SIPP Users' Guide* discusses specific steps involved in transforming the data. It also includes sample SAS code.

Analysts should note that edit and imputation procedures differ for some variables. In addition, starting with the 1991 Panel, SIPP missing wave imputation procedures have created a situation in which data may be present in the full panel files but not in the core wave files.

Merging Two or More Topical Module Files

Analysts may wish to study the relationship between subject areas covered by different topical modules. For example, they might want to study the relationship between education and training history as reported in the second wave of the 1996 Panel and employment history as reported in the first wave of the 1996 Panel. In that case, they will need to link topical module files. In some panels, all of those data are reported in the same wave and no merge is necessary.



Topical module files are relatively simple to merge because they all have the same format (one record per person). Also, the ID variables are the same across files, except that the names for those variables differ between the 1996 and pre-1996 files (e.g., SSUID vs. ID). Nevertheless, analysts need to be cautious:

- Prior to the 1996 Panel, a variable name sometimes was used in different topical module files for different variables.

SIPP*tip*

Key variables have different names in the core wave and full panel files. Analysts should check the technical documentation to make sure that they are matching information as they intend.

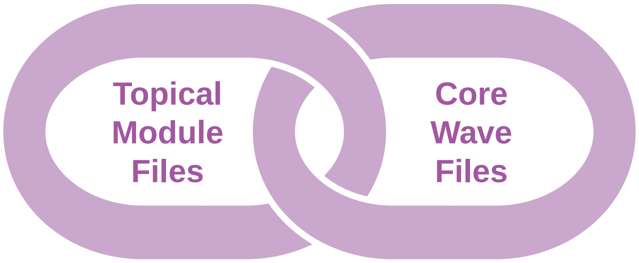
- Not all people with records in one topical module file will have records in another topical module file. Household composition may have changed from one wave to the next, and this will be reflected in the topical module files. In addition, nonmatches might occur because of nonresponse. Also, universes for topical modules may differ.
- The substantial number of nonmatches across topical modules complicates the choice of weights. Analysts might instead want to use one of the weights from the full panel files.

Analysts wishing to measure change with data from the topical module files should be careful because of changes in measurement over time.

In addition, apparent changes across pre-1996 topical modules could be due to real changes reported by the respondent or to edit and longitudinal inconsistencies.

Merging Topical Module and Core Wave Files

It is sometimes necessary to merge topical module files with data from the core wave files. Analysts should be careful when selecting which core wave file to use—some topical modules sought information about the interview month, while the core wave files contain information about a different reference month.



Topical module files have one record per person, while core wave files have as many as four. Therefore, three options exist for merging topical module and core wave files:

1. Select a single month from the core wave files.
2. Spread the topical module data across all records from the core wave file, which results in a final file in person-month format.

3. Create a single record for each person from the core wave file and merge the topical module data to that record.

Analysts should execute the following steps:

1. Create an extract from the core wave file.
2. Apply the appropriate algorithm, as shown in Chapter 13 of the *SIPP Users' Guide*.
3. Sort the core wave extract by using the sort keys that uniquely identify people in the core wave file.
4. Create an extract from the topical module, and sort.
5. Merge the core wave extract with the topical module extract and sort. Sort keys will be different for the 1996 Panel and previous panels. tip

Merging Topical Module and Full Panel Files

This procedure applies to panels prior to 1996. There are times when analysts will want to merge topical module and full panel files. For example, if the full panel weights are needed for the planned analysis, they must come from the full panel files. tip

The full panel files contain a record for every person who was ever a member of a SIPP household. Therefore, every person with a record in a topical module file should have a record in the full panel file. Analysts working with a person-month file may nonetheless find nonmatches.

For this type of linkage, analysts should carry out the following steps:

1. Create an extract from the full panel file.
2. If the person-month format is desired, apply the appropriate algorithm (see Chapter 13 of the *SIPP Users' Guide*), but rename the ID variables to match those used in the topical module files.
3. Sort the full panel extract.

SIPP*tip*

In the pre-1996 Panels, there will likely be nonmatches between the file types because people who were present in the interview month (topical module files) may not have been present during any of the previous 4 months (core wave files).

tip

The edit and imputation procedures used with the full panel files are believed to introduce less error than the procedures used with the core wave files. Thus, when the same core items are available from the core wave and full panel files, analysts may prefer to use the full panel files.

4. Create an extract from the desired topical module file, and sort.
5. Merge the two extracts by using the appropriate ID variables.

Nonmatches and Other Anomalies

SIPP follows a group of people over a period of time. Original sample members are followed throughout the time period unless they die or leave the sample universe by moving to an ineligible location, such as a nursing home, a military barracks, or another country. Secondary sample members are part of SIPP only when they live with an original sample member.

Nonmatches occur when analysts merge across waves for any file types. Respondents may be in one data file and not another for a number of reasons:

- Original sample members move to (or back from) ineligible locations or drop out of the sample but not the sample universe.
- Secondary sample members move into or out of the sample.
- The person is a newborn.
- Missing wave data imputed in the full panel file is not in the core wave or topical module files.
- The person was in a merged household and received new ID information.

Entering and Exiting the Population

There is a fundamental distinction between situations in which people leave the sample because they leave the SIPP sample universe and situations in which they leave the sample but are still part of the population.

In general, when nonmatches occur because of people entering or exiting the population of the sample, data should not be imputed and weights should not be adjusted for the period of their absence.

Analysts can employ a number of strategies to deal with these nonmatches:

- They can drop leavers from the sample entirely and not adjust the weights of the retained cases. The remaining sample now represents the population that existed at both Time 1 and Time 2. tip
- Event-history models can also be used, with a person's exit from the population as one of the competing outcomes.

Sample Attrition

Sample attrition occurs when people leave the sample but remain part of the population represented by the sample. Several options exist for handling such cases. Analysts can choose to:

- Impute the missing data
- Eliminate cases with missing data and poststratify the weights for the retained cases
- Use a subset of cases with complete data and Census Bureau–provided weights
- Use other missing data methods to provide estimates and standard errors tip

Missing Wave Imputation

Beginning with the 1991 Panel, the Census Bureau has applied a missing wave imputation procedure to full panel files. Persons with missing data for one wave but complete data for two adjacent waves have data imputed.

If these cases were person-level nonrespondents who had data imputed with different methods in the core wave files, the data in their full panel and core wave records will differ. Other persons may have data for the missing wave only in the full panel file. For a complete explanation of the handling of missing wave data in SIPP, refer to the study "Compensating for Missing Wave Data in the Survey of

SIPP*tip*

Dropping leavers from the sample is simple to do, but analysts then cannot draw inferences about the part of the population that has left. For example, the economic profiles of people leaving the sample to enter prison or a nursing home will likely differ from the profiles of those who remain in the sample.

tip

All of the methods for handling sample attrition require caution. Chapter 13 of the SIPP Users' Guide *presents an in-depth discussion of the possible pitfalls.*

Income and Program Participation" by Williams and Bailey, which can be accessed from the SIPP home page under Publications.

The correct procedure for dealing with these nonmatches depends on which weights will be used.

- If weights come from the core wave or topical module files, analysts should drop observations from the full panel files that are not present in the cross-sectional files.
- If weights come from the full panel file, the Census Bureau suggests using the procedures for sample attrition.

Merged Households

Nonmatches can occur when the Census Bureau changes ID numbers for sample members. In panels before 1996, there were two very rare occasions when this happened. The first was when two separate sampling units with original sample members merged together, perhaps because of a marriage. The Census Bureau changed the identification information of one set of original sample members to agree with the other set.

The second instance occurred when a SIPP household split into new households, gained new secondary sample members in each, and later recombined with the secondary sample members coming along. In the recombined household, the secondary sample members from one of the earlier split households were assigned new person numbers.

Different file types recorded this information differently. Chapter 13 of the *SIPP Users*' *Guide* discusses this situation in-depth and tells how analysts can search the core wave file for these people. Analysts can then change the identification information, duplicate and merge the records, or treat the person with the new identity as two people, as is done in the full panel files.